Fuzzy Inf. Eng. (2017) 9: 345–357

Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Information and Engineering
http://www.elsevier.com/locate/fiae

ORIGINAL ARTICLE

Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset

Mehrbakhsh Nilashi · Othman Ibrahim · Mohammad Dalvi
Hossein Ahmadi · Leila Shahmoradi

Received: 2 September, 2016/ Revised: 12 May, 2017/
Accepted: 27 August, 2017/

Abstract As a chronic disease, diabetes mellitus has emerged as a worldwide epidemic. Providing diagnostic aid for diabetes disease by using a set of data that contains only medical information obtained without advanced medical equipment, can help numbers of people who want to discover the disease or the risk of disease at an early stage. This can possibly make a huge positive impact on a lot of peoples lives. The aim of this study is to classify diabetes disease by developing an intelligence system using machine learning techniques. Our method is developed through clustering, noise removal and classification approaches. Accordingly, we use SOM, PCA and NN for clustering, noise removal and classification tasks, respectively. Experimen-

Corresponding Author: Mehrbakhsh Nilashi (✉)
Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
email: nilashidotnet@hotmail.com
Othman Ibrahim (✉)
Department of Computer Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran
email: othmanibrahim@utm.my
Mohammad Dalvi (✉)
Department of Mechatronics Engineering, University of Isfahan, Isfahan, Iran
email: mohammaddalvi@hotmail.com
Corresponding Author: Hossein Ahmadi (✉)
Health Information Management Department, 5th Floor, School of Allied Medical Sciences, Tehran University of Medical Sciences, No. 17, Farredanesh Alley, Ghods St, Enghelab Ave, Iran
email: hosseinis3007@gmail.com
Corresponding Author: Leila Shahmoradi (✉)
Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran
email: lshahmoradi@tums.ac.ir
Peer review under responsibility of Fuzzy Information and Engineering Branch of the Operations Research Society of China.

http://dx.doi.org/10.1016/j.fiae.2017.09.006

tal results on Pima Indian Diabetes dataset show that proposed method remarkably improves the accuracy of prediction in relation to methods developed in the previous studies. The hybrid intelligent system can assist medical practitioners in the healthcare practice as a decision support system.

Keywords Diabetes disease diagnosis · Clustering · PCA · Neural Network

1. Introduction

Diabetes has been one of the leading health problems in the United States [37]. It has attained the dubious distinction of becoming the fifth leading cause of disease-related death [19]. Diabetes is a chronic endocrine disorder affecting the bodys metabolism and resulting in structural changes affecting the organs of the vascular system [10, 13]. Generally, diabetes is characterized as existing in two major forms: (a) insulin-dependent (Type I) and (b) noninsulin-dependent (Type II). The latter appears to be the more common, accounting for 80% of all diabetic cases [19]. The Pima are one of the most studied populations regarding diabetes, not only among American Indians, but in the world [24]. The most studied populations regarding diabetes is Pima, not only among American Indians, but in the world [24]. The samples of studied populations regarding diabetes refer to discrete Type-2 positive and negative instances. The only way for the diabetes patient to live with this disease is to keep the blood sugar as normal as possible without serious high or low blood sugars this is achieved when the patient uses a correct management (therapy) which may include diet and exercising, taking oral diabetes medication or using some form of insulin [19]. On the other hand treating the diabetes disease is also a difficult, an expensive and a complex task for the medical staff [19]. There are number of important things to record about the patient and disease that help the doctors to make an optimal decision about the patient to make his/her life better.

Machine learning deals with the development of technologies which allow machines to learn. The challenge is to create algorithms that can take a group of patterns (on a broader range, the existing knowledge) and automatically make new inferences from the initial information, with or without human intervention. From the machine learning perspective, classification is the problem of identifying a set of observations into several categories, basing on the training result of a subset of observations whose belonging category is known. The unsupervised learning is defined as cluster analysis. It is also called clustering. Clustering is a process of putting a set of observations into several reasonable groups according to certain measure of similarity within each group. The clustering problem has been addressed in many diseases diagnosis systems [20, 8, 40, 33, 34, 35]. This reflects its broad appeal and usefulness as one of the steps in exploratory health data analysis.

There is a vast sea of different techniques and algorithms used in data mining especially for supervised machine learning techniques; therefore, selecting the appropriate techniques has been a challenge among researchers in developing the diabetes disease diagnosis systems [15, 22]. Hence, in order to improve predictive accuracy

of diabetes disease classification, a new method is proposed by applying noise removal, classification and clustering techniques. To the best knowledge of the authors, there is no implementation of classification method (NN), clustering method (SOM) and noise removal method (PCA) for diabetes disease diagnosis form the real-world dataset.

Our study at hand is organized as follows: In Section 2 we present the related-work. In Section 3 the research methodology and all techniques incorporated to the proposed method are explained. In Section 4, the evaluations of methods are presented. Finally, we conclude our work in Section 5.

2. Related Work

Polat et al. [41] used discriminant analysis and the Support Vector Machine (SVM) for diabetes classification. Using 10-fold cross validation, they achieved 82.05% of accuracy on Pima diabetes dataset. Kayaer and Yldrm [23] developed a method using the General Regression Neural Network (GRNN) for diabetes classification. The method was tested on PID and achieved 80.21% accuracy for classification. Aslam et al. [1] proposed a method using the Genetic Programming (GP) for diabetes classification. The method includes three stages: features selection, features generation and testing. Two classifiers, the k-Nearest Neighbor (k-NN) and the SVM, were used for evaluating the selected features. The authors tested the performance of method using Pima Indians diabetes dataset. A hybrid intelligent system was developed by [22] using the Fuzzy Neural Network (FNN) and the Artificial Neural Network (ANN). They evaluated the method on two public medical datasets, Pima Indians diabetes and Cleveland heart disease. Using k-fold cross-validation, the method obtained classification accuracies of 84.24% and 86.8% for Pima Indians diabetes dataset and Cleveland heart disease dataset, respectively. An intelligent system was proposed by [15] for diagnosis of diabetes. The method was based on the Small-World Feed Forward ANN (SW-FFANN). The accuracy of the method was 91.66%. Ganji and Abadeh [16] developed a method,the FCS-ANTMINER, by the Ant Colony Optimization (ACO). They extracted a set of fuzzy rules to classify the diabetes disease. The obtained classification accuracy was 84.24%. An intelligent diagnosis system, the LDA-ANFIS, was developed by Dogantekin et al. [12] for diabetes using the Linear Discriminant Analysis (LDA) classification method and the Neuro-Fuzzy (ANFIS) system. The classification accuracy of the LDA-ANFIS was about 84.61%. A comparative study of diabetes disease on Pima Indian diabetes disease was conducted by [44]. They used multilayer NN which was trained by the Levenberg-Marquardt (LM) algorithm and probabilistic NN. An automatic diagnosis system, the LDACMWSVM, was developed for diabetes by [4]. They used the Morlet Wavelet Support Vector Machine (MWSVM) Classifier and the Linear Discriminant Analysis (LDA). Their method classification accuracy was about 89.74%. From the literature on diabetes disease diagnosis form experiments with Long Beach and Cleveland Clinic Foundation, we found that at the moment there are no implementations of the PCA, Gaussian mixture model with the Self-Organizing Map (SOM), and the NN method for distinguishing between presence and absence of diabetes disease in patients. This research

accordingly tries to develop a diabetes disease diagnosis intelligent system based on these methods. Overall, in comparison with research efforts found in the literature, in this research:

- SOM is used for data clustering. The clustering problem has been addressed in many diseases diagnosis systems [20, 8, 40]. This reflects its broad appeal and usefulness as one of the steps in exploratory health data analysis. In this study, SOM clustering is used as an unsupervised classification method to cluster the data of experimental dataset into similar groups.
- NN is used for data classification. NN is widely employed in diagnosis of diseases for their efficiency and robustness. It is a promising classification approach which has been used in many researches on diseases classification [23, 15].
- PCA is used for dimensionality reduction and dealing with the multi-collinearity problem in the experimental data. This technique has been used in developing in many disease diagnosis systems to eliminate the redundant information in the original health data [9, 4].

By combination of SOM, PCA, NN, a hybrid intelligent system is proposed to increase the predictive accuracy of diabetes disease.

3. Methodology

Focusing on the prediction and classification of diseases, the present study uses PCA, SOM, and classification (NN) methods. The general framework of proposed model is shown in Fig. 1. We propose to rely on classification methods to learn the classification functions. Additionally, PCA is employed for dimensionality reduction and to overcome the multi-collinearity problem of the datasets. These methodologies are addressed in the following sections.

Dataset. The Pima aboriginals diabetes dataset is provided at the courtesy of National Institute of Diabetes and Digestive and Kidney Diseases and Vincent Sigillito of the Applied Physics Laboratory of the Johns Hopkins University who was the original donor of the dataset. The actual data itself is obtained by the author of this research from the website of the UCI (University of California at Irvine) [3]. This data has been used in the past by the researchers to investigate possible vital signs that may be used to indicate the presence of diabetes within patients according to World Health Organization (WHO) standards. There are a total of 768 training instances included in this dataset. Each training instance has 8 features and class variable that provides the label for that training instance (see Table 1). The features are Number of times pregnant, Plasma glucose concentration, Diabetes pedigree function, Triceps skin fold thickness (mm), Diastolic blood pressure (mmHg), 2-h serum insulin (mU/ml), Body mass index (kg/m2) and Years of age. The class variable takes on the binary value of 0 or 1 with 0 indicating a healthy person and 1 indicating a diabetic patient.

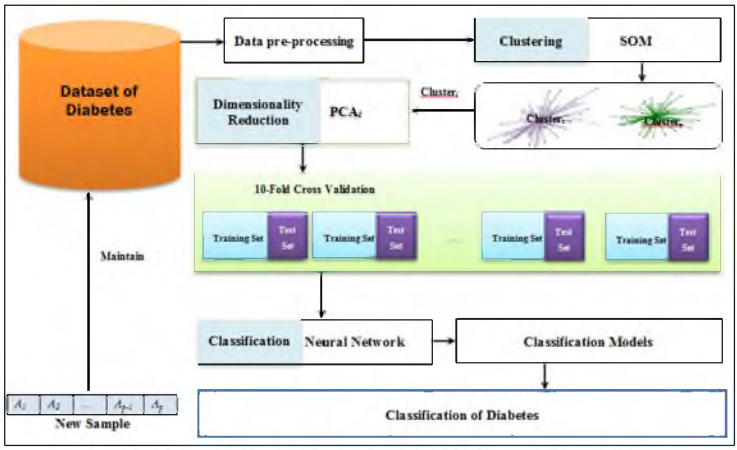

Fig. 1: Proposed method for the diabetes diseases diagnosis

Table 1: Description of the features of Pima Indian diabetes dataset.

Variable	Feature Label	Variable Type	Range
X1	Number of times pregnant	Integer	0-17
X2	Plasma glucose concentration in a 2 hour oral glucose tolerance test	Real	0-199
X3	Diastolic blood pressure	Real	0-122
X4	Triceps skin fold thickness	Real	0-99
X5	2 hour serum insulin	Real	0-846
X6	Body mass index	Real	0-67.1
X7	Diabetes pedigree function	Real	0.078-2.42
X8	Age	Real	21-81
Y	Class	Integer	0,1

Self-organizing Map Clustering. SOM has been proved useful as data clustering and projection tools. It uses a small set of well-organized neurons to represent a data set and creates a neighborhood preserving mapping from a high-dimensional data space onto a low-dimensional grid of neurons. These neurons are located on a low-dimensional, usually 2-dimensional, grid which provides us a convenient surface to project and summarize data. Interestingly, the mapping roughly preserves the most important topological and metric relationships of data, and thus, inherent clusters in data. The main advantage of clustering using SOM is retention of the underlying structure of the input space, while the dimensionality of the space is reduced. Kohonens original SOM algorithm can be seen to define a special recursive regression

process, in which only a subset of the models are processed at every step. The SOM establishes a projection through an iterative learning procedure. It behaves like a exible net that folds onto the cloud formed by the input data during the learning. In each learning step, one data item is chosen randomly, and the distances between it and all the weight vectors of the output neurons are calculated using a certain distance metric (e.g., the Euclidean distance). A The SOM depicted in Fig. 2 in maps from the input data space Rn onto a two-dimensional array of nodes is called a lattice.

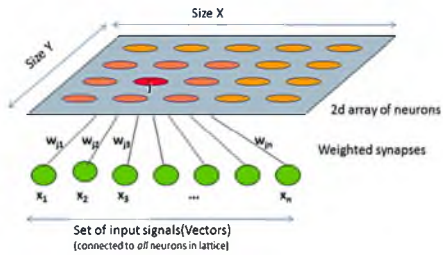

Fig. 2: A SOM of size X × Y

Principal Component Analysis. PCA is a statistical technique for multivariate analysis and is used as a dimensionality reduction technique in data compression to retain the essential information and are easy to display [29· 32· 36]. The method identifies patterns in data and represents the data in a way that highlights similarities and differences. The central idea is to reduce the number of dimensions of the data while preserving as much as possible of the variations in the original dataset. PCA has four goals. The first goal is to extract the most information from the data. The second goal is to compress the data by only keeping the most characterizing information. The third goal is to simplify the description of the data and the fourth goal is to enable analysis of the structure of the observations. The analysis enables conclusions to be drawn regarding the used variables and their relations. The analysis is performed through transforming the data to a new set of variables, called the Principal Components (PCs). The PCs are uncorrelated and ordered so that the first few PCs retain most of the variations of the total dataset. The first PC describes the dimension in which the data has the biggest variation (variance) and the second component describes the dimension in which it has the second largest variation (variance). PCA is chosen for this study because the method exemplifies a category of analysis methods. If the data has linear relations and is correlated, as data often is in medical datasets, the method will give a compression that maintains a high amount of the information in the original dataset. The described solution saves a compact summary of the data, which is derived by applying ideas from statistics to enable an analysis while preserving its characteristics. In this research, we also apply PCA on the clusters generated by SOM to transform the correlated variables in the original dataset into a set of variables which are linearly uncorrelated.

Neural Network. ANN is one of the most popular approaches used extensively in machine learning, which is involved in the development of algorithms that enable computers to learn [45]. Similar to the human brain, an ANN also consists of a number of neurons. These neurons are interconnected processors. They are connected by weighted links used to pass signals between the neurons. Learning ability is accomplished by repeated adjustments of these weights until errors on the network output as compared to the training data are minimized. A NN can be considered as a form of non-linear mapping from several input variables to several output variables. Several parameters are set up to support the mapping. In order to constitute a NN, three major types of layers has to be set up; i.e. input, hidden, and output layers [43]. Each layer contains a number of neurons. There are various kinds of NN architectures. The architectures are differed in terms of the number of layers, the number of neurons, and the connections between the neurons. They also have different learning algorithms. In general, a supervised NN can be used to solve two kinds of problem, which are classification and regression problems. In classification problem, each output obtained from a NN is assigned to one of a number of classes whereas outputs of the regression problem represent continuous values. In this research, the problem of classification is addressed. As different NNs produce different predictions and uncertainties values, hence multiple networks have been utilized in this research. In this research, feedforward backpropagation NNs are used for the classification.

The selection of this network architecture is based on its popularity and efficiency. The model used in this research has three layer. The relationship between the input variables ($x_1, x_2, \cdots, x_n$) and output variable (y) has the following mathematical formula:

In the above formula, m and n are number of input nodes and hidden nodes, respectively. $w_{i,j}$ ($i = 0, 1, 2, \cdots, n$; $j = 0, 1, 2, \cdots, m$) and w_j ($j = 0, 1, 2, \cdots, q$) denotes model parameters (connection weights). For training of NN, different back-propagation techniques are used. Resilient back-propagation, Conjugate gradient back-propagation and LM algorithm are some of the techniques which are used for training. In this research, we used NN with resilient back-propagation training algorithm

$$y_t = w_0 + \sum_{i=1}^{m} w_j.g(w_0 j + \sum_{i=1}^{n} w_i(i, j).x_i(t, j)). \tag{1}$$

Cross-Validation. Cross-validation is a statistical method that in this research is used for the performance evaluation of learning algorithms and performance of a predictive model on an unknown dataset. For this reason, using cross-validation, the datasets used in the research are divided into several equally sized subsets (see Fig. 3). The learning model is then trained on some subsets known as training sets. After training process, the model is tested on the remaining subsets, known as test sets. According to the number of subsets partitioned, researcher tests k-fold cross-validation. For 10-fold cross-validation, researchers use 10 result of 10-fold cross-validation. In the experiments of this research, for the training of models, it is considered to test different 10 for 10-fold cross-validation, so that researchers can make sure that there

are enough training instances to learn the models [25].

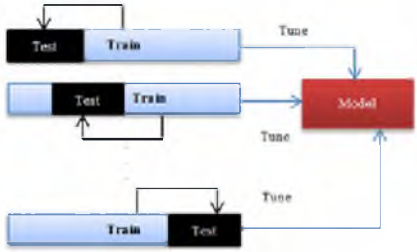

Fig. 3: K-Fold cross validation

4. Results of Methods

The experimental results of the proposed method on real-world datasets are explained in this section. Here, the results of applying all incorporated methods in the proposed system are discussed.

Clustering with SOM Algorithm. In this research, SOM algorithm is applied on experimental dataset. We considered SOM 2×2,SOM 2×3,SOM 3×3,SOM 3×4,SOM 4×4 and SOM 4×5 for SOM to see which type obtains the best clustering accuracy and observed that with learning rate=0.5, SOM 2×3 (6 clusters) can provide the best clustering quality. In Fig.4, the clusters generated by all types of SOM are visualized by projecting the observations in the first two dimensions generated by PCA.

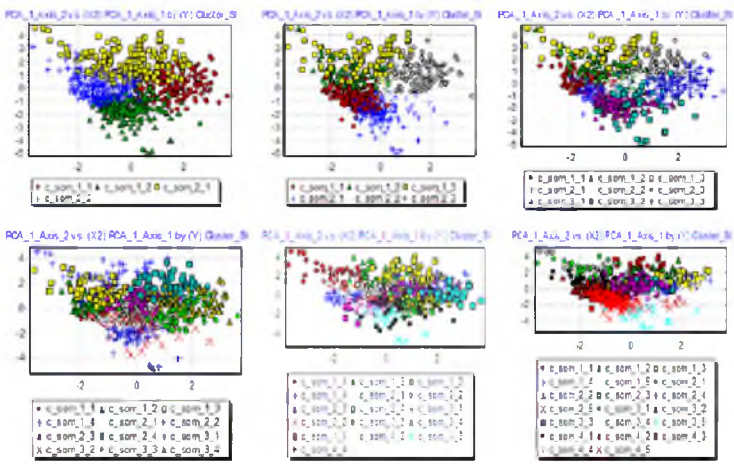

Fig. 4: Clusters generated by SOM

PCA Evaluation. As PCA generates PCs instead of original factors, choosing the right number selection of these PCA is an important task. If we select too many factors, we include noise from the sampling fluctuations in the analysis. If we choose too few factors, we lose relevant information, the analysis is incomplete. As we know that the eigenvalue associated to a factor corresponds to its variance. Thus, the eigenvalue indicates the importance of the factor. The higher is the value, the higher is the importance of the factor. The eigenvalues for each factor can be indicators for its importance. In this study, we have applied the rule proposed by [6]. Accordingly, we create scree plots, that show the eigenvalues of the factors. In the scree plots, we can simply detect elbows to decide the number of PCA to be used in the classifcation process.

We employed the PCA technique for the clusters of experimental dataset obtained by SOM algorithm. Based on rule proposed by [6], in PID, for Cluster 1, we included the elbow into the selection i.e., we selected $k = 2$ factors. Indeed, the eigenvalues associated with the 2nd factor was high. In addition, three PCs for Clusters 2 and 4 and four PCs for Clusters 3, five and six were chosen.

Performance Evaluation of NN. This section provides the experimental results of diabetes disease classification with NN classifier applied on PID. In addition, comparison experiments with other methods in the literature are performed using on the same dataset. The models of classification were trained under a 4 GHz processor PC and Microsoft Windows 7 running MATLAB 7.10 (R2010a). We applied NN on experimental dataset clustered by SOM algorithm. To show the predictive accuracy of the proposed methods, we use AUC (Area Under Curve) of ROC (Receiver Operating Characteristic) chart. ROC is a graphical display that provide the measure of classification accuracy of the model using sensitivity and specificity [50]. For predicting events, sensitivity in ROC can be used as a measure of accuracy which can be calculated by dividing the true positive over total actual positive. For predicting nonevents, specificity is can be used as a measure of accuracy which can be calculated by dividing true negative over the total actual negative of a classifier for a range of cutoffs. For NN classification, 70% of data is considered for training, 20% for validating and 10% for testing the model. Learning rate and number of hidden layer are set to 0.5 and 8, respectively. In addition, the model is trained for 200 epochs (see Fig. 5). For original dataset, the weights from inputs to hidden layer and hidden to outputs layer with 8 neurons are presented in Tables 2 and 3.

In order to experimentally demonstrate the effectiveness of PCA, SOM clustering and NN, we conduct the experiments on the public PID and compare with the methods of developed in the previous works for accuracy. The average accuracy obtained by the proposed method is about 92.28% for all clusters. We compare the accuracy of our proposed method with the classification accuracy of the methods General Regression Neural Network [23], GDA-LSSVM [41], MWSVM [5] and SW-FFANN [15] for PID. The performance of the classifiers that were compared with our method is shown in Table 4. From the results shown in this table, our proposed method proves to have a better accuracy (92.28%) in relation to the other classification systems. Compared to General Regression Neural Network (80.21%), GDA-LSSVM

Table 2: The weights from hidden to outputs layer in NN.

Inputs	Neuron "1"	Neuron "2"	Neuron "3"	Neuron "4"	Neuron "5"	Neuron "6"	Neuron "7"	Neuron "8"
X1	−1.02	0.30	−0.03	0.01	0.17	−0.21	0.17	0.50
X2	−2.44	−0.92	−0.9	−0.98	−0.95	−0.94	−0.96	−1.64
X3	−0.30	0.14	0.51	0.16	0.15	0.42	0.15	1.29
X4	1.06	0.17	0.24	−0.03	0.07	0.19	0.05	−0.14
X5	1.71	−0.56	0.24	−0.52	−0.54	−0.23	−0.53	−0.17
X6	−0.08	−1.34	−0.66	−1.01	−1.18	−0.76	−1.15	0.34
X7	−0.19	−0.66	0.43	−0.63	−0.68	0.03	−0.64	−1.13
X8	−3.54	0.59	0.62	0.57	0.58	0.56	0.57	−3.11
bias	−1.42	−0.46	0.67	−0.10	−0.32	0.48	−0.26	−0.41

Table 3: The weights from inputs to hidden layer in NN.

Neuron	Healthy	Diabetic
Neuron "1"	−1.9254	1.9254
Neuron "2"	−0.8376	0.8379
Neuron "3"	−0.6384	0.6384
Neuron "4"	−0.7011	0.7008
Neuron "5"	−0.7774	0.7773
Neuron "6"	−0.5581	0.5581
Neuron "7"	−0.7559	0.756
Neuron "8"	−1.4903	1.4904
bias	2.8759	−2.8759

(79.16%), MWSVM (89.74%) and SW-FFANN (91.66%), our classification, clustering and noise removal techniques help to improve the classification accuracy of diabetes disease by more than 12%, 13%, 2% and 0.6%, respectively. This shows the effectiveness of incorporating the clustering and PCA techniques for the classification accuracy of diabetes disease.

Table 4: Comparison of proposed method with other classifiers for PID.

Method	Reference	Accuracy
General Regression Neural Network	[23]	80.21%
GDA-LSSVM	[41]	79.16%
MWSVM	[5]	89.74%
SW-FFANN	[15]	91.66%
PCA-SOM-NN	This Study	92.28%

5. Conclusion and Future Work

In this paper, we propose a new hybrid intelligent system for diabetes disease classification using machine learning techniques. We applied SOM clustering algorithm

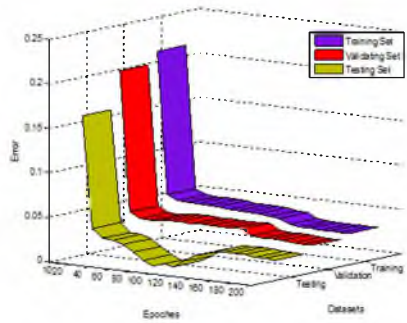

Fig. 5: NN and the classification errors

to cluster the experimental diabetes disease dataset and NN for classification of disease types. In addition, PCA was used for dimensionality reduction and to address multi-collinearity in the dataset. In order to analyze the effectiveness of the proposed method and validate the system, several experiments were conducted on PID. The dataset was taken from UCI. The results indicated that the method which combines clustering, PCA and NN is used to obtain good classification accuracy. All of the approaches, used in this study, may also be applicable to other classification problems within the medical domain. However, there is still plenty of work in conducting researches on clustering, noise removal and classification methods to diabetes disease diagnosis in order to exploit all their potential and usefulness. In the future work, more attention should be paid to datasets for disease classification by using the incremental machine learning approaches. Hence, in our future study, we plan to evaluate the proposed method on additional datasets and in particular on large datasets to show the effectiveness of incremental methods to computation time of large data in relation of the non-incremental ones.

Acknowledgements

Appreciation goes to the anonymous reviewers whose comments helped us to improve the manuscript.

References

[1] M.W. Aslam, Z. Zhu, A.K. Nandi, Feature generation using genetic programming with comparative partner selection for diabetes classification, Expert Systems with Applications 40 (2013) 5402-5412.

[2] M. Awad, Y. Motai, J. N?ppi, H. Yoshida, A clinical decision support framework for incremental polyps classification in virtual colonoscopy, Algorithms 3 (2010) 1-20.

[3] K. Bache, M. Lichman, UCI Machine Learning Repository. 2013.

[4] D. Çalişir, E. Dogantekin, A new intelligent hepatitis diagnosis system: PCACLSSVM, Expert Systems with Applications 38 (2011) 10705-10708.

[5] D. ?ali?ir, E. Do?antekin, An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier, Expert Systems with Applications 38 (2011) 8311-8315.

[6] R.B. Cattell, The scree test for the nnmber of factors, Multivariate behavioral research 1 (1966) 245-276.

[7] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM transactions on intelli-

gent systems and technology (TIST) 2 (2011) 27.

[8] C.H. Chen, A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection, Applied Soft Computing 20 (2014) 4-14.

[9] H.L. Chen, C.C. Huang, X.G. Yu, X. Xu, X. Sun, G. Wang, S.J. Wang, An efficient diagnosis system for detection of Parkinsons disease using fuzzy k-nearest neighbor approach, Expert Systems with Applications 40 (2013) 263-271.

[10] S. Court, E. Sein, C. McCowen, A. Hackett, J. Parkin, Children with diabetes mellitus: perception of their behavioural problems by parents and teachers, Early Human Development 16 (1988) 245-252.

[11] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B (methodological) (1977) 1-38.

[12] E. Dogantekin, A. Dogantekin, D. Avci, L. Avci, An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS, Digital Signal Processing 20 (2010) 1248-1255.

[13] [13] L.E. Egede, Diabetes, major depression, and functional disability among US adults, Diabetes care 27 (2004) 421-428.

[14] L.E. Egede, Y. Miohel, Perceived difficulty of diabetes treatment in primary care: does it differ by patient ethnicity? The Diabetes Educator 27 (2001) 678-684.

[15] O. Erkaymaz, M. Ozer, Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes, Chaos, Solitons and Fractals 83 (2016) 178-185.

[16] M.F. Ganji, M.S. Abadeh, A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis, Expert Systems with Applications 38 (2011) 14650-14659.

[17] V. Gerl, L. Lhotska, M. Murgas, V.D. Radisavljevic, V. Krajca, V. Kremen, An Incremental Approach to Clinical EEG Data Classification, 6th European Conference of the International Federation for Medical and Biological Engineering, Springer, 2015, pp. 489-492.

[18] P.M. Hall, A.D. Marshall, R.R. Martin, Incremental eigenanalysis for classification, BMVC, 1998, pp. 286-295.

[19] B.A. Hamburg, G.E. Inoff, Relationships between behavioral factors and diabetic control in children and adolescents: a camp study, Psychosomatic Medicine 44 (1982) 321-339.

[20] E.R. Hruschka, N.F. Ebecken, Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach, Neurocomputing 70 (2006) 384-397.

[21] Y.G. Jung, M.S. Kang, J. Heo, Clustering performance comparison using K-means and expectation maximization algorithms, Biotechnology and Biotechnological Equipment 28 (2014) S44-S48.

[22] H. Kahramanli, N. Allahverdi, Design of a hybrid system for the diabetes and heart diseases, Expert Systems with Applications 35 (2008) 82-89.

[23] K. Kayaer, T. Y?ld?r?m, Medical diagnosis on Pima Indian diabetes using general regression neural networks, Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), 2003, pp. 181-184.

[24] W.C. Knowler, D.J. Pettitt, P.H. Bennett, R.C. Williams, Diabetes mellitus in the Pima Indians: genetic and evolutionary considerations, American Journal of Physical Anthropology 62 (1983) 107-114.

[25] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Ijcai, Stanford, CA, 1995, pp. 1137-1145.

[26] S. Lekkas, L. Mikhailov, Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases, Artificial Intelligence in Medicine 50 (2010) 117-126.

[27] N.C. Long, P. Meesad, H. Unger, A highly accurate firefly based algorithm for heart disease prediction, Expert Systems with Applications 42 (2015) 8221-8231.

[28] J.F.G. Molina, L. Zheng, M. Sertdemir, D.J. Dinter, S. Sch?nberg, M. R?dle, Incremental learning with SVM for multimodal classification of prostatic adenocarcinoma, PloS One 9 (2014).

[29] B. Moore, Principal component analysis in linear systems: Controllability, observability, and model reduction, IEEE Transactions on Automatic Control 26 (1981) 17-32.

[30] G. Nathiya, S. Punitha, M. Punithavalli, An analytical study on behavior of clusters using k-means, em and k-means algorithm, International Journal of Computer Science and Information Security 7 (2010) 185-190.

[31] M. Nilashi, H. Ahmadi, L. Shahmoradi, M. Salahshour, O. Ibrahim, A soft computing method for

mesothelioma disease classification, Journal of soft Computing and Decision Support Systems 4 (2017) 16-18.

[32] M. Nilashi, O. bin Ibrahim, N. Ithnin, N.H. Sarmin, A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCACANFIS, Electronic Commerce Research and Applications 14 (2015) 542-562.

[33] M. Nilashi, O. Bin Ibrahim, A. Mardani, A. Ahani, A. Jusoh, A soft computing approach for diabetes disease classification, Health Informatics Journal (2016).

[34] M. Nilashi, O. Ibrahim, A. Ahani, Accuracy improvement for predicting Parkinsons disease progression, Scientific Reports 6 (2016).

[35] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, A knowledge-based system for breast cancer classification using fuzzy logic method, Telematics and Informatics 34 (2017) 133-144.

[36] M. Nilashi, O.B. Ibrahim, N. Ithnin, R. Zakaria, A multi-criteria recommendation system using dimensionality reduction and Neuro-Fuzzy techniques, Soft Computing 19 (2015) 3173-3207.

[37] A.A. Onitilo, R.V. Stankowski, R.L. Berg, J.M. Engel, G.M. Williams, S.A. Doi, A novel method for studying the temporal relationship between type 2 diabetes mellitus and cancer using the electronic medical record, BMC Medical Informatics and Decision Making 14 (2014).

[38] C. Ordonez, E. Omiecinski, FREM: fast and robust EM clustering for large data sets, Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM, 2002, pp. 590-599.

[39] D. Pelleg, A.W. Moore, X-means: Extending k-means with efficient estimation of the number of clusters, ICML, (2000) 727-734.

[40] K. Polat, Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering, International Journal of Systems Science 43 (2012) 597-609.

[41] K. Polat, S. Gne?, A. Arslan, A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine, Expert systems with Applications 34 (2008) 482-487.

[42] L.M. Saini, Peak load forecasting using Bayesian regularization, Resilient and adaptive backpropagation learning based artificial neural networks, Electric Power Systems Research 78 (2008) 1302-1310.

[43] L.M. Silva, J.M. de S, L.A. Alexandre, Data classification with multilayer perceptrons using a generalized error function, Neural Networks 21 (2008) 1302-1310.

[44] H. Temurtas, N. Yumusak, F. Temurtas, A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications 36 (2009) 8610-8615.

[45] G.S.M. Thakur, R. Bhattacharyya, S.S. Mondal, Artificial neural network based model for forecasting of inflation in India, Fuzzy Information and Engineering 8 (2016) 87-100.

[46] S. Tortajada, M. Robles, J.M. Garcła-Gmez, Incremental logistic regression for customizing automatic diagnostic models, Data Mining in Clinical Medicine (2015) 57-78.

[47] M. Vakili, M. Karami, S. Delfani, S. Khosrojerdi, Experimental investigation and modeling of thermal radiative properties of f-CNTs nanofluid by artificial neural network with LevenbergCMarquardt algorithm, International Communications in Heat and Mass Transfer 78 (2016) 224-230.

[48] C.H. Wang, C.H. Kao, W.H. Lee, A new interactive model for improving the learning performance of back propagation neural network, Automation in Construction 16 (2007) 745-758.

[49] C.J. Wu, On the convergence properties of the EM algorithm, The Annals of Statistics (1983) 95-103.

[50] W. Yu, T. Liu, R. Valdez, M. Gwinn, M.J. Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, BMC Medical Informatics and Decision Making 10 (2010).